IETE Journal of Research

ISSN: 0377-2063 (Print) 0974-780X (Online) Journal homepage: https://www.tandfonline.com/loi/tijr20

Human Face Detection and Tracking using Skin Color Modeling and Connected Component Operators

Prem Kuchi, Prasad Gabbur, P Subbanna Bhat & S Sumam David

To cite this article: Prem Kuchi, Prasad Gabbur, P Subbanna Bhat & S Sumam David (2002) Human Face Detection and Tracking using Skin Color Modeling and Connected Component Operators, IETE Journal of Research, 48:3-4, 289-293, DOI: 10.1080/03772063.2002.11416288

To link to this article: https://doi.org/10.1080/03772063.2002.11416288

IETE Journal of Research
Vol 48, Nos 3 & 4, May-August, 2002, pp 289-293

Human Face Detection and Tracking using Skin Color Modeling and Connected Component Operators

PREM KUCHI, PRASAD GABBUR, P SUBBANNA BHAT, AND SUMAM DAVID S, MIETE
Department of Electronics and Communication Engineering, Karnataka Regional Engineering College, Surathkal, Karnataka 574 157, India.

Face Recognition (FR) systems are increasingly gaining more importance. Face detection and tracking in a complex scene forms the first step in building a practical FR system. In this paper, a method to detect and track human faces in color image sequences is described. Skin color classification and morphological segmentation is used to detect face(s) in the first frame. These detected faces are tracked over subsequent frames by using the position of the faces in the first frame as the marker and detecting for skin in the localized region. Specific advantages of this approach are that skin color analysis method is simple and powerful, and the system can be used to detect/track multiple faces.

Indexing terms: Face detection/tracking, Skin color modeling, Connected component operators, Structuring element.

RECENT years have seen tremendous amount of research being carried out in the field of automatic face recognition. Automatic face recognition is a process of identifying a test face image with one of the faces stored in a prepared face database. Real world images need not necessarily contain isolated face(s) that can directly serve as inputs to a FR system. Hence, there is a need to isolate or segment facial regions to be fed to a FR system. Most of the time, a video sequence of the scene is available using which a person may have to be recognized. For recognition, we need the face position in which it is best recognizable by the present day FR algorithms. Hence, a robust system that detects and tracks a face is necessary. Face detection and tracking becomes an important task with the growing demand for content-based image functionality. Recent standardization efforts (MPEG 4/7) also point in this direction.

Though human beings detect/track faces with very little effort, it is not easy to train a computer to do so. In pattern recognition parlance, human face is a complex pattern. Different poses and gestures of the face accentuate complexity. The detection scheme must operate flexibly and reliably regardless of the lighting conditions, background clutter in the image, multiple faces in the image, as well as variations in face scale, pose and expression. The system should be able to detect the face even under small occlusions. Therefore, a systematic approach, keeping in mind both the robustness and the computational complexity of the algorithm is called for. Various methods have been proposed in the literature for face detection. Important techniques include template-matching [1], neural network [2], feature-based methods [3-7], motion based [4,8] and face space methods [9].

Paper No 91-C;

SKIN COLOR MODELING

The inspiration to use skin color analysis for initial classification of an image into probable face and non-face regions stems from a number of simple but powerful characteristics of skin color. Firstly, processing skin color is simpler than processing any other facial feature. Secondly, under certain lighting conditions, color is orientation invariant. The major difference between skin tones is intensity e.g., due to varying lighting conditions and different human race [10]. The color of human skin is different from the color of most other natural objects in the world. An attempt to build comprehensive skin and non-skin models has been done in [11].

One important factor that should be considered while building a statistical model for color is the choice of a Color Space. Segmentation of skin colored regions becomes robust only if the chrominance component is used in analysis. Therefore, we eliminate the variation of luminance component as much as possible by choosing the CbCr plane (chrominance) of the YCbCr color space to build the model. Another reason for the choice of YCbCr domain is its extensive use in digital video coding applications. Research has shown that skin color is clustered in a small region of the chrominance plane [11]. The distribution of the training skin pixels in the CbCr plane is given in Fig 1.

The figure shows that the color of human skin pixels is confined to a very small region in the chrominance space. Motivated by the results in the figure, the skin color

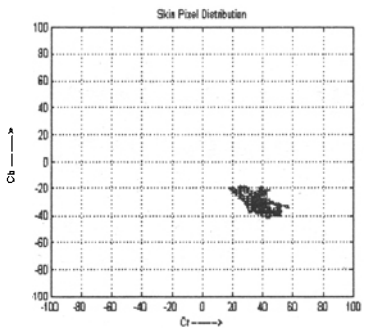


Fig 1 Skin pixel distribution

distribution in the chrominance plane is modeled as a unimodal Gaussian [10]. A large database of labeled skin pixels is used to build the Gaussian model. The mean and the covariance of the database characterize the model. Images containing human skin pixels as well as non-skin pixels are collected. The skin pixels from these images are carefully cropped out to form a set of training images.

Let $\mathbf{C} = [\mathrm{Cb\ Cr}]^T$ denote the chrominance vector of an input pixel. Then the probability that the given pixel lies in the skin distribution is given by

$$p(c/\mathrm{skin}) = \frac{\exp\left[-\frac{1}{2}(c-\mu_s)^T \Sigma_s^{-1}(c-\mu_s)\right]}{2\pi\sqrt{|\Sigma_s|}} \quad (1)$$

where μ_s and Σ_s represent the mean vector and the covariance matrix respectively of the training pixels. The above equation gives the probability of a pixel occurring given that it is a skin pixel. A similar model is developed for non-skin pixels using another database of pixels that do not represent skin.

Let *p(c/non-skin)* represent the conditional probability of the occurrence of a pixel, given that it lies in the non-skin region. Then the probability that a pixel represents skin given its chrominance vector *c*, *p(skin/c)* can be evaluated using Bayes' theorem. Since a pixel can be either a skin or a non-skin pixel but not both, and assuming that skin and non-skin pixels occur with equal probability, we assign probability of 0.5 for occurrence of both skin as well as non-skin pixels. Therefore,

$$p(skin/c) = \frac{p(c/skin)}{p(c/skin) + p(c/non-skin)} \quad (2)$$

Thus the problem reduces to the calculation of the two conditional probabilities and computing the above ratio to give the probability of a pixel being skin given its chrominance vector *c*. An input image is analyzed pixel-by-pixel evaluating the skin probability at each pixel position. This results in a gray level image where the gray value gives the probability of the pixel representing skin. This is called the Skin Probability image, ϕ, defined as

$$\phi(i,j) = a.p(skin/c_{ij}) \quad (3)$$

where *a* is a constant.

The skin probability image thus obtained is thresholded to obtain a binary image. Selection of an optimum threshold, *T*, is important as it affects the later stages of the detection process. A lower threshold is better because further analysis is based on connected component operators. Increasing the threshold will increase the chances of losing certain skin regions exposed to adverse lighting conditions, during thresholding. Also, the extra regions that get retained in the image because of the lower threshold can be removed using connected component operators. So, using a lower threshold will not compromise performance.

CONNECTED COMPONENT OPERATORS

Connected component operators are non-linear filters that eliminate parts of the image, while preserving the contours of the remaining parts. This simplification property makes them attractive for segmentation and pattern recognition applications.

Figure 2 shows a connected operator, Ψ, operating on a binary image *A*, consisting of two connected components. Ψ is such that it retains the shape of one component while completely removing the other component. In general, connected operators use certain decision criteria to either retain or eliminate a connected component without affecting the other components.

An opening by reconstruction operator is applied first on the binary image that is obtained after thresholding. This operation is nothing but erosion followed by dilation using a certain structuring element (SE). Erosion removes small and thin isolated noise-like components that have very low probability of representing a face. Dilation preserves those components that are not removed during erosion. Hence, the effect of using area open is removal of small and bright regions of the thresholded image. This is followed by closing by reconstruction. Here, dilation followed by erosion with a certain structuring element is performed. Initial dilation closes any small holes that may have been created during opening in probable face regions. Erosion removes the extra pixels that are added to the contour of the preserved components. During both opening and closing the size of the structuring element should not be more than that of the smallest face the system is designed to detect. The remaining connected components are isolated.

A set of shape based connected operators, Compactness, Solidity and Orientation, are applied over these remaining components to decide whether they

Fig 2 Example of a connected component operator

represent a face or not. These operators make use of basic assumptions about the shape of the face. Components that can be excluded from the face candidates based on their shape are removed. These simple but effective decision criteria rely on the combinations of the area, A, the perimeter, P, and the size, D_x and D_y of the min-max box of the connected component. Thus these features have to be computed only once for the three operators.

Compactness of a connected component is defined as the ratio of its area to the square of its perimeter.

$$\text{Compactness} = \frac{A}{P^2} \tag{4}$$

This criterion is maximized for circular objects. Faces are nearly circular in shape and hence face components exhibit a high value for this operator. A threshold is fixed for this operator based on the observations on various face components. If a particular component shows a compactness value greater than this threshold it is retained for further analysis, else discarded.

Solidity of a connected component is defined as the ratio of its area to the area of the min-max box (rectangular bounding box).

$$\text{Solidity} = \frac{A}{D_x D_y} \tag{5}$$

Solidity gives a measure of area occupancy of a connected component within its min-max box dimensions. The solidity also assumes a high value for face components. If the solidity of a component is lesser than a specified threshold value, it is eliminated, otherwise retained for further analysis.

Orientation is nothing but the aspect ratio of the min-max box surrounding the component.

$$\text{Orientation} = \frac{D_y}{D_x} \tag{6}$$

It is assumed that normally face components have orientation within a certain range. This range is found out based on observations on a number of images. If a component's orientation falls out of this range, the component is eliminated. A lot of non-face components that have solidity and compactness of a face component can be removed using orientation operator. For example, using orientation, a face component can be separated from an elongated pipe component, a horizontal elliptic component etc.

The remaining unwanted components are removed using Normalized Area. It is the ratio of the area of the connected component to that of the largest component that remains after the application of the above three operators. In images containing multiple faces it is assumed that the smallest face component has an area that is not less than a certain fraction of the largest face component. This is arrived at based on our observations of practical images containing multiple faces. The connected components that remain after the application of all the above operators contain faces.

TRACKING

To track face(s) in a given video sequence, the detection step described above is performed over one frame. The detected positions and the min-max boxes around face components serve as markers for locating face(s) in the next frame. To project into the next frame, we increase the dimensions of the min-max box, both horizontally and vertically, by 10% of D_x and D_y. Within this new min-max box region in the next frame, skin color analysis is done and a new min-max box for the skin colored region is computed. The newly computed min-max box dimensions serve as D_x and D_y for the next frame. In a scene, if the face is very near to the camera even a small amount of face movement causes a lot of motion from one frame to another. Where as, if the face is far from the camera, even considerable movement may not result in significant motion between adjacent frames. The adaptive increase in the min-max box dimensions while projecting into the next frame, depending on D_x and D_y values of the present frame, compensates for this differing amount of motion between frames when the face(s) in the scene are away or near to the camera.

We see that to locate face(s) in a video, the computationally expensive detection step need not be performed for every frame. Instead skin analysis and that too in localized regions is done to track faces. This reduces a lot of computational overhead and at the same time giving a robust tracking performance. This marker projection is repeated over several frames, N. Periodically the detection step is also performed so that the system does not miss new face(s) that may enter the scene. These steps of periodically detecting face(s) after fixed frame intervals followed by marker projection for the next several frames together track face(s) through a video sequence. The disadvantage of this tracking method is that, if a person enters and leaves a scene within this N number of frames, he/she is not detected/ tracked. So, the number, N, should be reasonably small.

IMPLEMENTATION & RESULTS

The first step to detect a human face against a complex background is to perform a skin color analysis of the image to isolate potential face regions. After calculating the probability of each pixel in the image to represent skin we threshold the image to segregate probable face candidates. The next step is to search for the face in the localized skin regions obtained in the above process. The connected components are isolated after performing area open/close by reconstruction. The shape and geometry based connected operators viz, Compactness, Solidity, Orientation and Normalized Area are finally applied to the

components to isolate face regions.

For skin color modeling we used a training set consisting of 10,000 skin pixels and 10,000 non-skin pixels. The skin pixel distribution obtained using this is shown in Fig 1. In order to arrive at threshold values for various connected operators we tested the operators on a number of face and non-face components. The following thresholds and constants were used:

Scaling constant, $a = 255$

Skin Probability binarization threshold, $T = 60$

Compactness threshold value = 0.025

Solidity threshold value = 0.5218

Orientation threshold range = 0.90 to 2.10

Normalized Area threshold = 0.35

Interval of performing detection, $N = 20$ frames

For tracking, once face detection is performed for a certain frame the detected positions are projected into subsequent frames followed by skin analysis within the projected regions to locate face(s) in these frames.

The algorithm has been implemented in Matlab 5.3. Typical detection results are shown in Fig 3. Tracking results for a few frames of a sample video sequence are shown in Fig 4. The success ratio is 0.82, when simulated on 120 general images down-loaded from the web, where the success ratio is defined as the ratio of the number of the faces detected to the number of faces input. Some failures were due to the non-inclusion of certain skin colored pixels in the model. Other failures were due to very small size of the face components that were eroded during opening by reconstruction performed using a single structuring element for all images irrespective of face sizes.

CONCLUSION

An algorithm has been developed to detect and track human face(s) in a color image sequence. The algorithm starts with human skin color modeling and uses it in isolating skin pixels (probable face regions). Skin color is found to be a powerful feature for isolating potential face candidates. It is also useful for detecting multiple human faces in an image. It is orientation independent.

(*a*) (*b*) (*c*)

Fig 3 Face detection results

(*a*) (*b*) (*c*)

Fig 4 Face tracking results

Connected Component Operators are applied on the thresholded skin probability image to isolate the final face components. The combination of the six operators used proved to be very effective. Skin color analysis followed by the use of shape based Connected Operators makes the system invariant to change in scale. For a higher detection performance, the structuring element used during open/close operations must be changed adaptively.

For tracking, a simple but effective approach has been followed. This involves projecting the face regions in the present frame as markers to the next frame, and detecting for skin in the localized regions. This reduces a lot of computational overhead by avoiding face detection in every frame. And since the min-max box dimensions to be projected into the next frame are increased adaptively based on the face size in the present frame, the tracking step is also scale independent.

Building a more robust skin model using larger number of skin and non-skin pixels would enhance the performance of the detector. Skin classification based on neural networks seems to be a promising method.

REFERENCES

1. Chellapa, Wilson, Sirohey, Human & Machine Recognition of Faces: A Survey, *Proceedings of the IEEE*, vol 83, no 5, pp 705-740, May 1995.

2. H A Rowley, S Baluja & T Kanade, Human face detection in visual scenes, CMU-CS-95-158R, Carnegie Mellon Univ, Nov,1995.

3. Ying Dai & Yasuaki Nakano, Face-texture model based on SGLD and its application in face detection in a color scene, *Pattern Recognition*, vol 29 (6), pp 1007-1017, 1996.

4. Choong Hwan Lee, Jun Sung Kim & Kyu Ho Park, Automatic human face location in a complex background using motion and color information, *Pattern Recognition*, vol 29 (11), pp 1877-1889, 1998.

5. Venu Govindaraju, Locating human faces in photographs, *International Journal of Computer Vision*, vol 19 (2), pp 129-146, 1996.

6. Qian Chen, Haiyuan Wu & Masahiko Yachida, Face detection by fuzzy pattern matching, *Proc IEEE International Conference on Computer Vision*, pp 591-596, 1995.

7. Young Ho Kwon & Niels da Victoria Lobo, Face detection using templates, *Proc International Conference on Pattern Recognition*, pp 764-767, 1994.

8. Makoto Kosugi, Human face search and location in a scene by multi-pyramid architecture for personal identification,

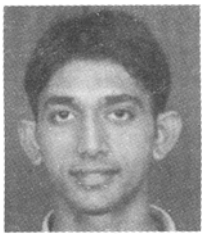
Systems and Computers in Japan, vol 26(6), pp 27-38, 1995.

9. M Turk & A P Pentland, Face recognition using eigenface, *Proc CVPR*, pp 586-593, 1991.

10. Menser, Wien, Segmentation & Tracking of Facial Regions in Color Image Sequences, RWTH, Aachen, Germany, 1999.

11. Jones, Rehg, Statistical Color Models with Application to Skin Detection, *Tech-Rep* CRL 98/11, Compaq Cambridge Research Lab 1998.

AUTHORS

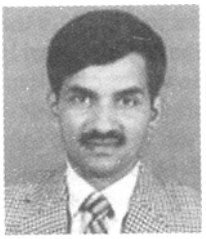


Prem Kuchi received his BE degree from Karnataka Regional Engineering College (KREC), Surathkal in 2001. He is presently pursuing his MS degree in the Dept. of Electrical Engineering at the Arizona State University, Tempe, and is also working as a research assistant in the Visual Computing and Communications Lab, ASU. He was a recipient of the Summer Research Fellowship awarded by the Jawaharlal Nehru Center for Advanced Scientific Research during Jul.-Sep. 2000. His present research interests include Image Processing, Pattern Recognition and Biometrics.

* * *

Prasad Gabbur received his BE degree from KREC, Surathkal in 2001. He is presently pursuing his MS degree in the Dept. of Electrical and Computer Engineering at the University of Arizona, Tucson and is a teaching assistant in the department. He was a Young Engineering research fellow of the Indian Institute of Science, Bangalore during Jul.-Sep. 2000. His research interests are in the fields of Signal and Image Processing, Pattern Recognition, and Computer Vision.

* * *



P Subbanna Bhat obtained his BE and MTech degrees from KREC Surathkal in 1974 and 1977 respectively, and the PhD degree in Power Electronics from the Department of Electrical Engineering, IIT Kanpur in 1984. He is presently Professor and Head of the Dept. of E&C, KREC Surathkal. He was a Visiting Researcher at Dept. of Electrical Engineering, Huddersfield University, UK during 1995. His research interests are in the areas of Digital Signal Processing, Aritificial Neural Networks and Fuzzy logic.

* * *

Sumam David S obtained her BTech degree from University of Kerala in 1985, and the MTech and PhD degrees from Department of Electrical Engineering, IIT Madras in 1986 and 1992 respectively. She is presently an Assistant Professor in the Dept of E&C, KREC Surathkal. She was a Visiting Researcher at Dept. of Computation, UMIST, Manchester, UK during 1995-96. She is a recipient of AICTE Career award for young teachers in 1998. Her research interests are in the areas of Audio-visual signal processing and VLSI for Signal processing.

* * *